# Digital Art History & the Web

## An Introduction to Web Archive Data Analysis

ARLIS/NA | May 21, 2025

bit.ly/arlis-data ⬇️

# Facilitators

**Karl Blumenthal**

Senior Web Archivist
Internet Archive

**Sumitra Duncan**

Head, Web Archiving Program
NYARC

**Julieann Mercer**

Program Assistant
Internet Archive

# Archiving & Data Services

ARCHIVE-IT

VAULT

ARCH

INTERNET ARCHIVE SCHOLAR

COMMUNITY WEBS

**1,300+ partners in 40+ countries.**

Projects and services include:

- Archive-It - partner driven web archiving
- Vault - digital preservation
- ARCH - text and data mining
- Scholar - >35m open research articles
- Digitization - books, manuscripts, A/V
- Community Programs - convening, capacity building and more!

# Thank you to the Kress Foundation!

## Digital Art History Grants

The Digital Art History Grants program is intended to foster new forms of research and collaboration as well as new approaches to teaching and learning. Support may also be offered for the digitization of important visual resources (especially essential art history photographic archives) in the area of pre-modern European art history; of primary textual sources (especially the literary and documentary sources of European art history); for promising initiatives in online publishing; and for innovative experiments in the field of digital art history.

# Agenda

1. Introduction, learning objectives
2. Archiving and Data Services, incl. CARTA
3. Web archives as data
4. Q&A: Web archives & WARC files
5. ARCH overview and data analysis demo
6. Discussion

# Learning objectives

- Learn web archive primary sources, research use cases and how to support them in a way that promotes computational engagement, scholarship, and research
- Understand the WARC file format for preservation and computational access
- Explore representative processes for analyzing computational research datasets with FOSS

# Collaborative ART Archive

Samuel H. Kress Foundation

NEH 1965 DEMOCRACY DEMANDS WISDOM

INSTITUTE of Museum and Library SERVICES

**Community formation and growth**
Member recruitment, collaborative meetings and discussions

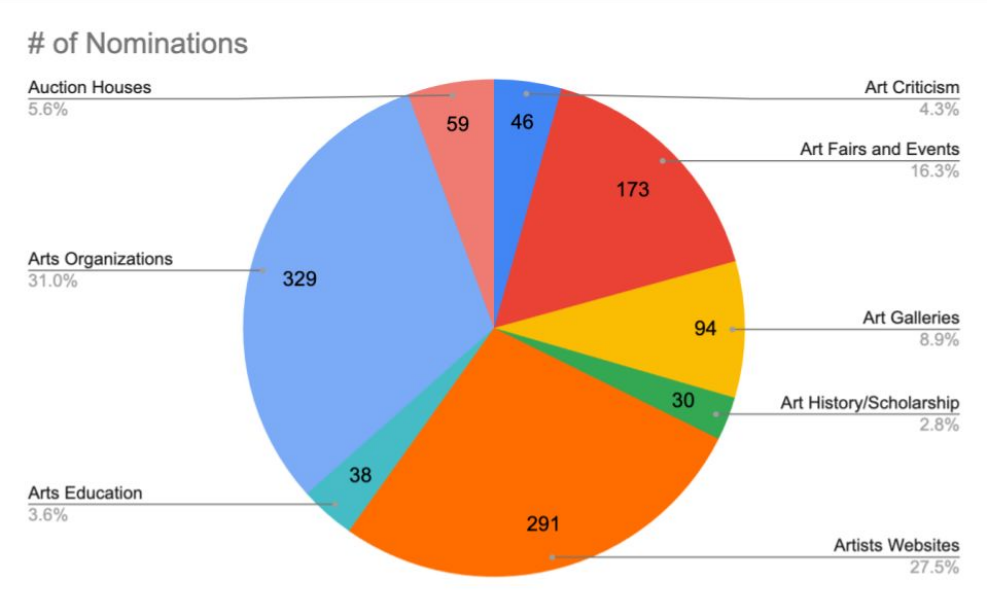
**Collection building**
Creating collections, preserving content, adding metadata

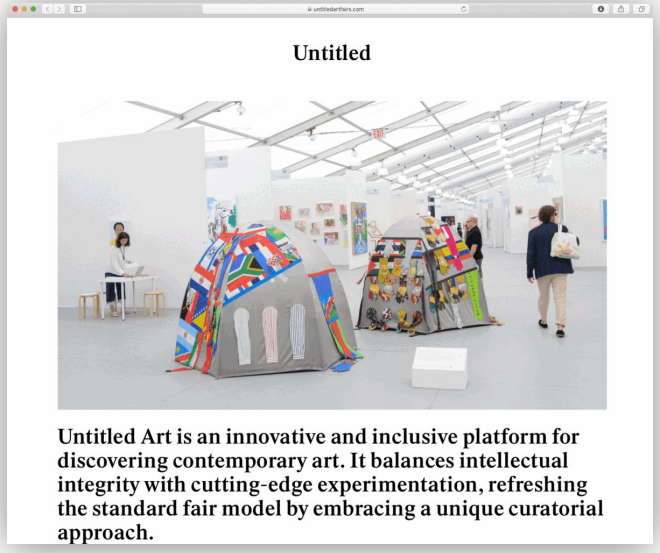
**Facilitating access and research**
CARTA access portal (https://carta.archive-it.org/), datathons and promotion of use with end users/researchers

# CARTA Participants

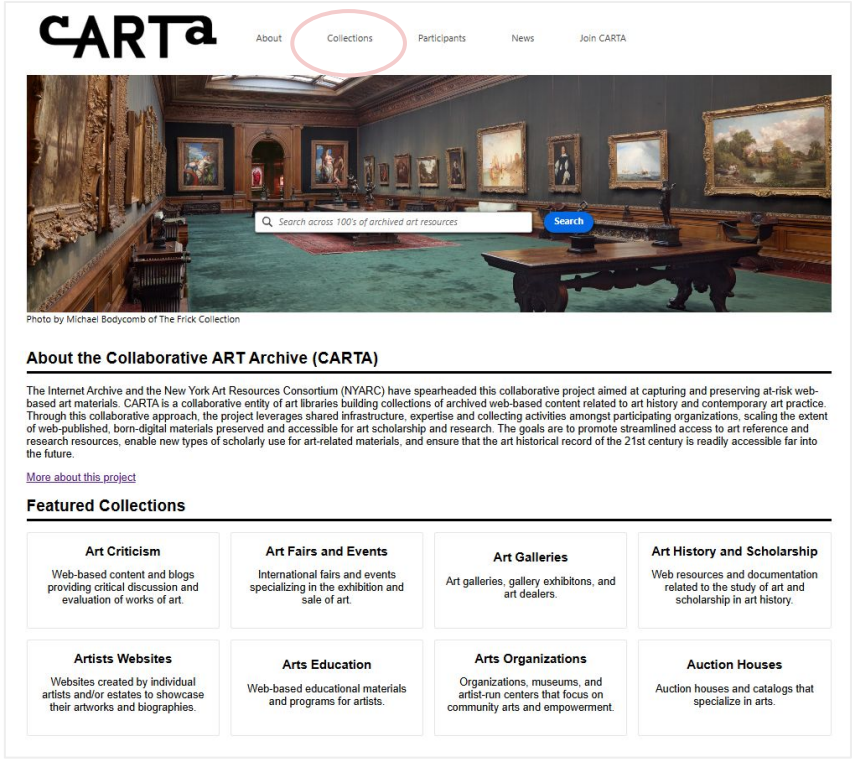
American Craft Council
American Folk Art Museum
Amon Carter Museum of American Art
ART | library deco
Art Gallery of Ontario
Art Institute of Chicago
Artexte
Bibliothèque et Archives nationales du Québec (BAnQ)
Birmingham Museum of Art
Cleveland Museum of Art
Detroit Institute of Arts
Emily Carr University of Art + Design
Fashion Institute of Technology - SUNY
Getty Research Institute (Getty Library)
Harvard University - Fine Arts Library
Harvard University - Graduate School of Design
Indianapolis Museum of Art at Newfields
Leonardo / ISAST
Los Angeles County Museum of Art (LACMA)
Maryland Institute College of Art (MICA)
Museum of Contemporary Art Chicago
Museum of Contemporary Art of Georgia (MOCA GA)
Museum of Glass
National Gallery of Art Library
National Gallery of Canada
New York Art Resources Consortium (NYARC)
Oklahoma State University
Philadelphia Museum of Art
Rhode Island School of Design
Rungh Cultural Society
San Francisco Museum of Modern Art (SF MOMA)
Smithsonian Libraries and Archives
Smithsonian Institution - Archives of American Art
Sterling and Francine Clark Art Institute Library
The Clyfford Still Museum
The Corning Museum of Glass
The Menil Collection
The Metropolitan Museum of Art
The Nelson-Atkins Museum of Art, Spencer Reference Library
University of Hawai'i at Mānoa Library
University of North Carolina, Greensboro
Walker Art Center

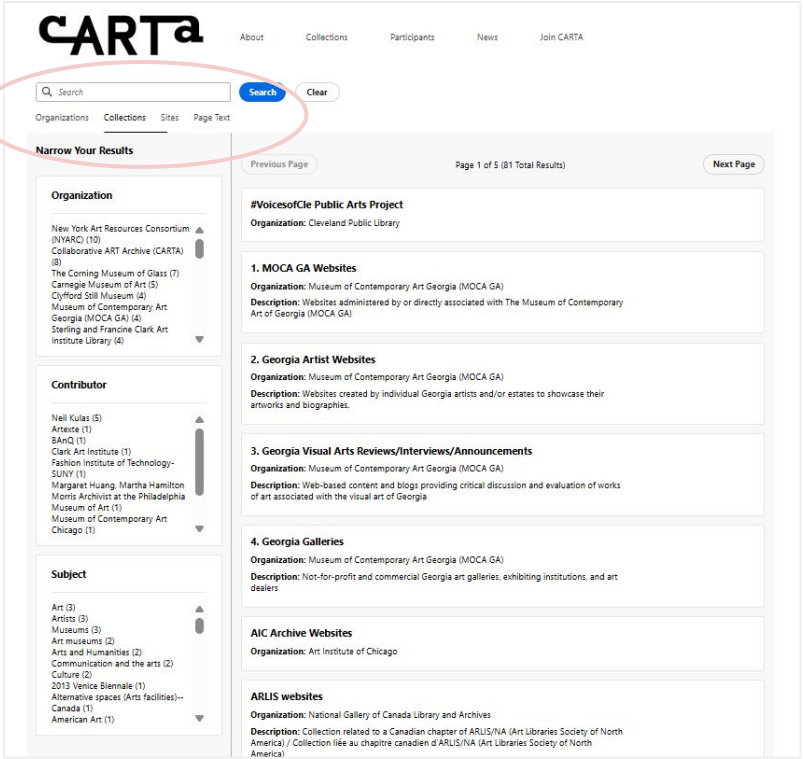
# CARTA's Collection Scope





Untitled
Untitled Art is an innovative and inclusive platform for discovering contemporary art. It balances intellectual integrity with cutting-edge experimentation, refreshing the standard fair model by embracing a unique curatorial approach.

# CARTA Access Portal


About
Collections
Participants
News
Join CARTA
Search across 100's of archived art resources
Search
Photo by Michael Bodycomb of The Frick Collection
About the Collaborative ART Archive (CARTA)
The Internet Archive and the New York Art Resources Consortium (NYARC) have spearheaded this collaborative project aimed at capturing and preserving at-risk web-based art materials. CARTA is a collaborative entity of art libraries building collections of archived web-based content related to art history and contemporary art practice. Through this collaborative approach, the project leverages shared infrastructure, expertise and collecting activities amongst participating organizations, scaling the extent of web-published, born-digital materials preserved and accessible for art scholarship and research. The goals are to promote streamlined access to art reference and research resources, enable new types of scholarly use for art-related materials, and ensure that the art historical record of the 21st century is readily accessible far into the future.
More about this project
Featured Collections
Art Criticism
Web-based content and blogs providing critical discussion and evaluation of works of art.
Art Fairs and Events
International fairs and events specializing in the exhibition and sale of art.
Art Galleries
Art galleries, gallery exhibitons, and art dealers.
Art History and Scholarship
Web resources and documentation related to the study of art and scholarship in art history.
Artists Websites
Websites created by individual artists and/or estates to showcase their artworks and biographies.
Arts Education
Web-based educational materials and programs for artists.
Arts Organizations
Organizations, museums, and artist-run centers that focus on community arts and empowerment.
Auction Houses
Auction houses and catalogs that specialize in arts.


About
Collections
Participants
News
Join CARTA
Search
Clear
Organizations
Collections
Sites
Page Text
Narrow Your Results
Previous Page
Page 1 of 5 (81 Total Results)
Next Page
Organization
New York Art Resources Consortium (NYARC) (10)
Collaborative ART Archive (CARTA) (8)
The Corning Museum of Glass (7)
Carnegie Museum of Art (5)
Clyfford Still Museum (4)
Museum of Contemporary Art Georgia (MOCA GA) (4)
Sterling and Francine Clark Art Institute Library (4)
Contributor
Neil Kulas (5)
Artexte (1)
BAnQ (1)
Clark Art Institute (1)
Fashion Institute of Technology--SUNY (1)
Margaret Huang, Martha Hamilton Morris Archivist at the Philadelphia Museum of Art (1)
Museum of Contemporary Art Chicago (1)
Subject
Art (3)
Artists (3)
Museums (3)
Art museums (2)
Arts and Humanities (2)
Communication and the arts (2)
Culture (2)
2013 Venice Biennale (1)
Alternative spaces (Arts facilities)--Canada (1)
American Art (1)
#VoicesofCle Public Arts Project
Organization: Cleveland Public Library
1. MOCA GA Websites
Organization: Museum of Contemporary Art Georgia (MOCA GA)
Description: Websites administered by or directly associated with The Museum of Contemporary Art of Georgia (MOCA GA)
2. Georgia Artist Websites
Organization: Museum of Contemporary Art Georgia (MOCA GA)
Description: Websites created by individual Georgia artists and/or estates to showcase their artworks and biographies.
3. Georgia Visual Arts Reviews/Interviews/Announcements
Organization: Museum of Contemporary Art Georgia (MOCA GA)
Description: Web-based content and blogs providing critical discussion and evaluation of works of art associated with the visual art of Georgia
4. Georgia Galleries
Organization: Museum of Contemporary Art Georgia (MOCA GA)
Description: Not-for-profit and commercial Georgia art galleries, exhibiting institutions, and art dealers
AIC Archive Websites
Organization: Art Institute of Chicago
ARLIS websites
Organization: National Gallery of Canada Library and Archives
Description: Collection related to a Canadian chapter of ARLIS/NA (Art Libraries Society of North America) / Collection liée au chapitre canadien d'ARLIS/NA (Art Libraries Society of North America)

CARTA Website/Portal: [https://carta.archive-it.org/](https://carta.archive-it.org/)

# Web archives as data

ARLIS/NA | May 21, 2025

bit.ly/arlis-data ⬇️

# Collections as data

> Increasingly, content providers in libraries seek to take library collections and present them as datasets that can be acted upon in various ways by machines or through use of computer-aided tools and methods. “Collections as data” is the idea that collections…can be used as data for researchers to analyze using computers.

Lampert, Cory & Emily Lapworth (2020) ”What do we Mean by ‘Collections as Data’ (CAD)?”

See also Thomas Padilla et al., *Collections as Data: Part to Whole*.

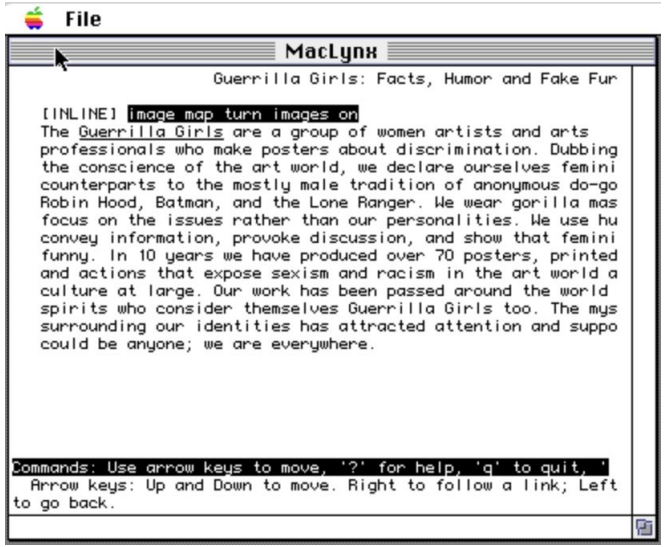
# Collections as data

Digital or computational research methods such as text mining, data visualization, mapping, image analysis, audio analysis, and network analysis automate steps in the research process that would take humans many hours to do, or are even impossible for humans to do manually.

Lampert, Cory & Emily Lapworth (2020) "What do we Mean by 'Collections as Data' (CAD)?"

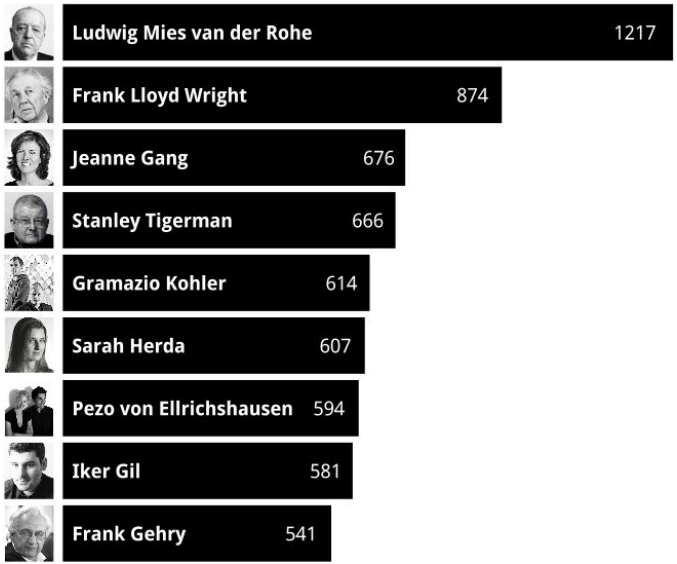
See also Thomas Padilla et al., *Collections as Data: Part to Whole*.

# What is a web archive?


File
MacLynx
Guerrilla Girls: Facts, Humor and Fake Fur
[INLINE] image map turn images on
The Guerrilla Girls are a group of women artists and arts professionals who make posters about discrimination. Dubbing the conscience of the art world, we declare ourselves femini counterparts to the mostly male tradition of anonymous do-go Robin Hood, Batman, and the Lone Ranger. We wear gorilla mas focus on the issues rather than our personalities. We use hu convey information, provoke discussion, and show that femini funny. In 10 years we have produced over 70 posters, printed and actions that expose sexism and racism in the art world a culture at large. Our work has been passed around the world spirits who consider themselves Guerrilla Girls too. The mys surrounding our identities has attracted attention and suppo could be anyone; we are everywhere.
Commands: Use arrow keys to move, '?' for help, 'q' to quit, '
Arrow keys: Up and Down to move. Right to follow a link; Left to go back.

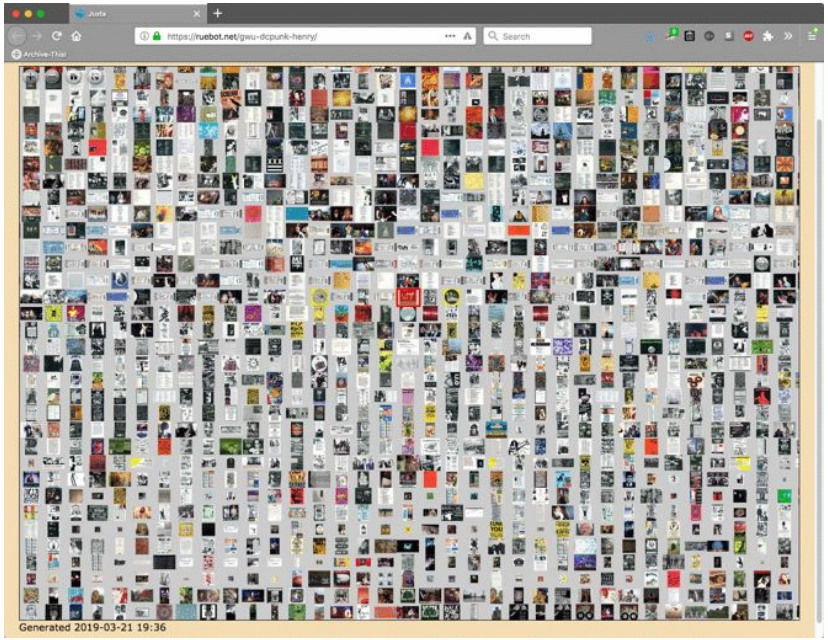
**A web archive is a collection of documents,** published originally to the worldwide web, preserved in their original context, and typically grouped by theme, event, subject area, or web domain.

# Web archives as data



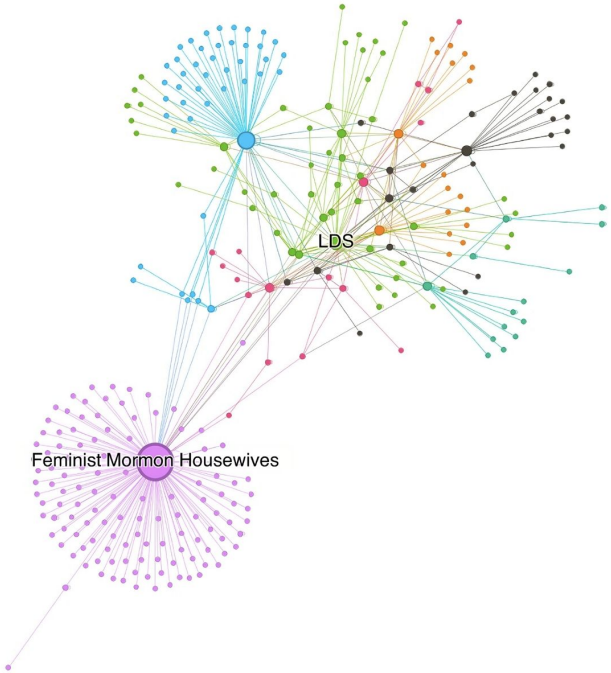

**10 most frequently mentioned persons**
in the Chicago Architecture Biennial 2015 web archive

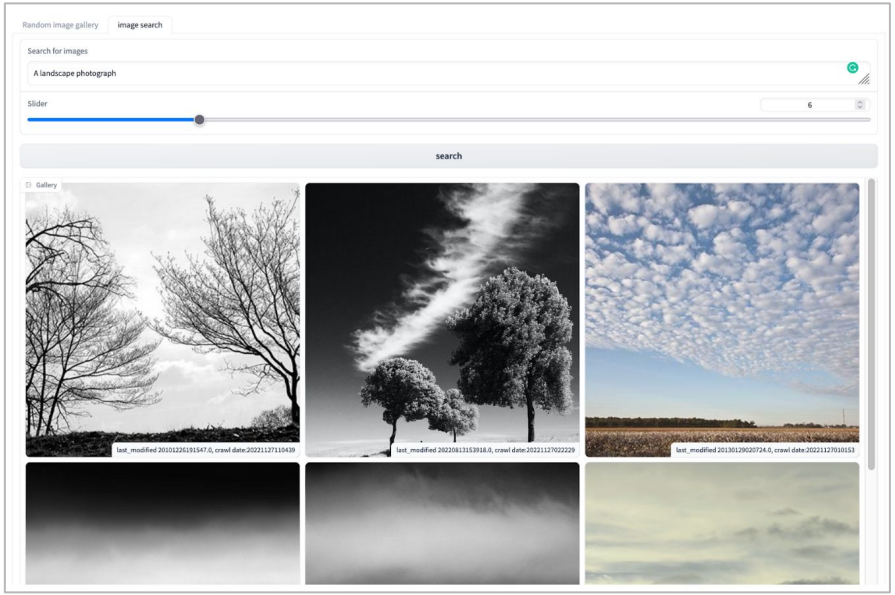
[Reading design from a distance: Mining enviornmental design web archives (2016)]




[Non-textual content in the DC Punk web archive (2021)]

# Web archives as data

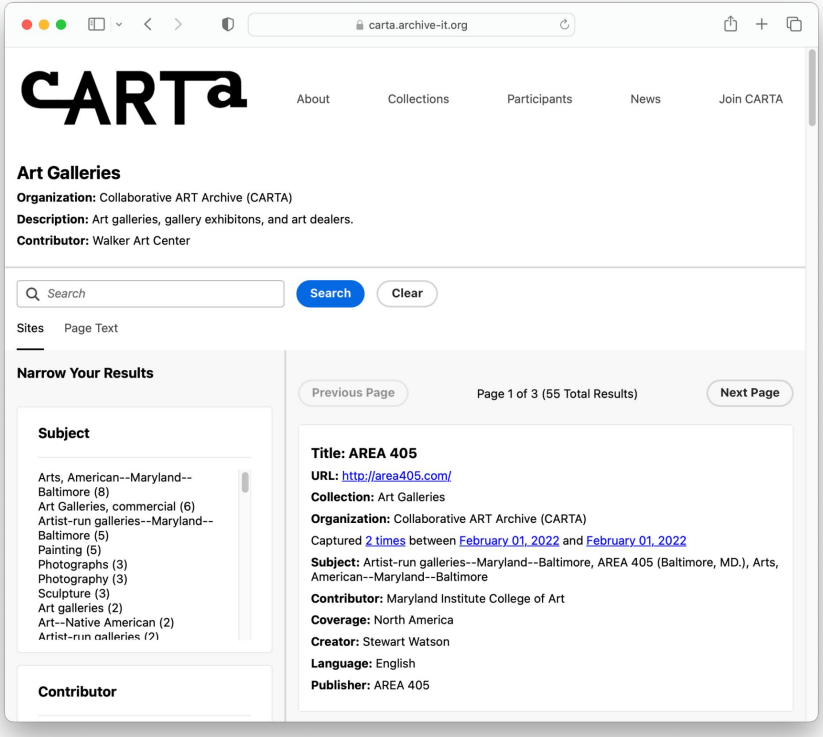


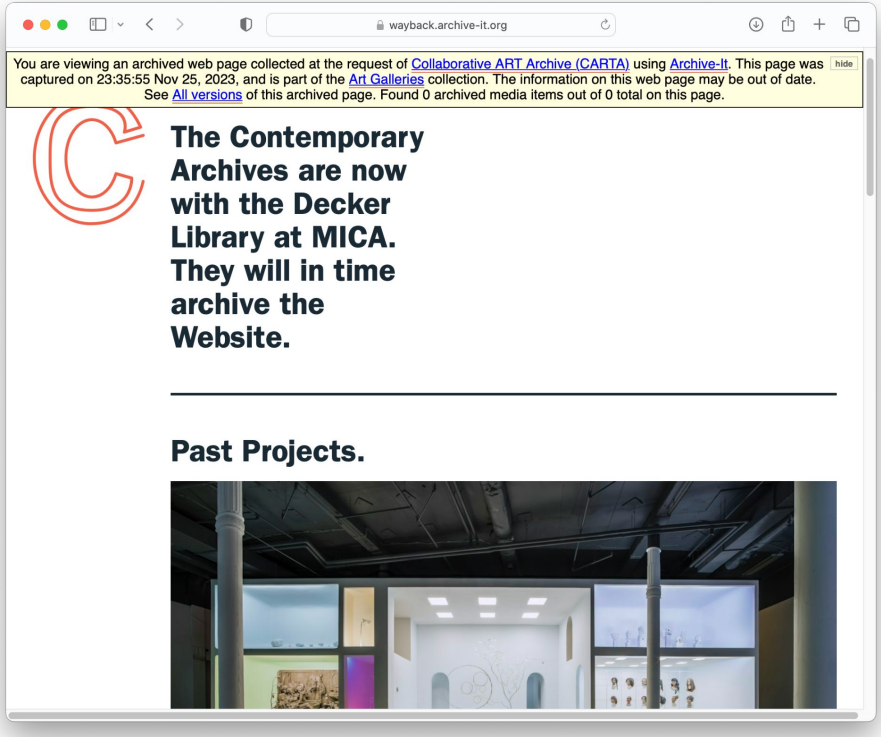




[Digital pioneers: Mormon mommy bloggers and building the "Bloggernacle" (2025)]

[Getting Started with Machine Learning and GLAM (Galleries, Libraries, Archives, Museums) Collections (2023)]

# Web archives as data

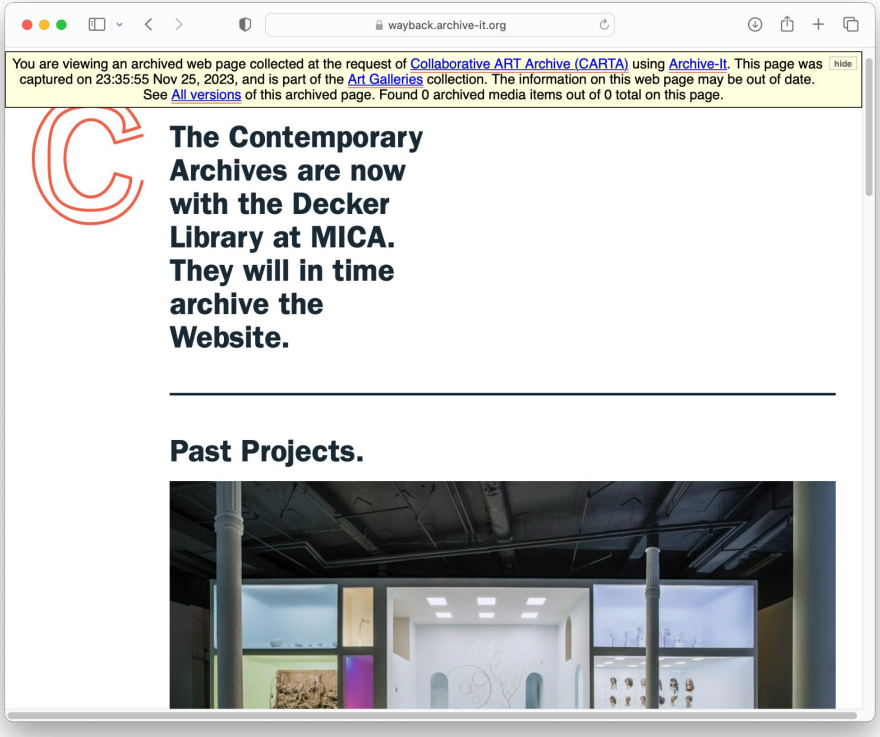




wayback.archive-it.org
You are viewing an archived web page collected at the request of Collaborative ART Archive (CARTA) using Archive-It. This page was captured on 23:35:55 Nov 25, 2023, and is part of the Art Galleries collection. The information on this web page may be out of date.
See All versions of this archived page. Found 0 archived media items out of 0 total on this page.
hide
The Contemporary Archives are now with the Decker Library at MICA. They will in time archive the Website.
Past Projects.

[Collaborative ART Archive (CARTA)]

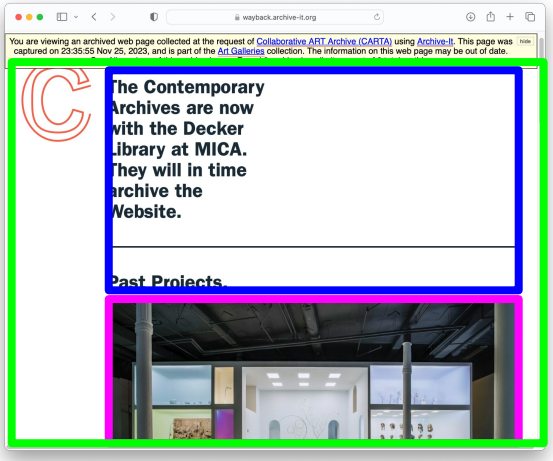
# Web archives as data

```
WARC/1.0
WARC-Type: response
WARC-Target-URI: https://www.contemporary.org/
WARC-Date: 2023-11-25T23:35:55Z
WARC-Payload-Digest: sha1:MLCZXFBA3KL244XXDHEWQQKBY3FW6QEN
WARC-Record-ID: <urn:uuid:e629a74e-9824-4684-8831-5b8571443386>
Content-Type: application/http; msgtype=response
Content-Length: 36264

HTTP/1.1 200 OK
Server: cloudflare
Date: Sat, 25 Nov 2023 23:35:55 GMT
Content-Type: text/html; charset=UTF-8
Connection: close
Vary: Accept-Encoding
Last-Modified: Wed, 22 Jan 2022 02:14:20 GMT

<!DOCTYPE html>
<html lang="en">
<head>
<meta charset="UTF-8">
<meta name="viewport" content="width=device-width, initial-scale=1">
href="https://contemporary.org/xmlrpc.php">
<title>The Contemporary, Baltimore, MD.</title>
```


wayback.archive-it.org
You are viewing an archived web page collected at the request of Collaborative ART Archive (CARTA) using Archive-It. This page was captured on 23:35:55 Nov 25, 2023, and is part of the Art Galleries collection. The information on this web page may be out of date.
See All versions of this archived page. Found 0 archived media items out of 0 total on this page.
hide
The Contemporary Archives are now with the Decker Library at MICA. They will in time archive the Website.
Past Projects.

# Web archives as data

Live website


wayback.archive-it.org
You are viewing an archived web page collected at the request of Collaborative ART Archive (CARTA) using Archive-It. This page was captured on 23:35:55 Nov 25, 2023, and is part of the Art Galleries collection. The information on this web page may be out of date.
The Contemporary Archives are now with the Decker Library at MICA. They will in time archive the Website.
Past Projects

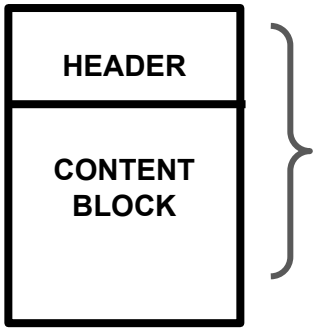
CSS

JPEG

HTML

WARC file

WARC record




```
WARC/1.0
WARC-Type: response
WARC-Target-URI: https://area405.com/2020/02/13/cdg-collection/
WARC-Date: 2020-03-19T18:57:00Z
WARC-Payload-Digest: sha1:MLCZXFBA3KL244XXDHEWQQKBY3FW6QEN
WARC-Record-ID: <urn:uuid:e629a74e-9824-4684-8831-5b8571443386>
Content-Type: application/http; msgtype=response
Content-Length: 72066

HTTP/1.1 200 OK
Server: nginx
Date: Thu, 19 Mar 2020 18:57:00 GMT
Content-Type: text/html; charset=UTF-8
Connection: close
Vary: Accept-Encoding
Last-Modified: Thu, 19 Mar 2020 18:55:56 GMT

<!DOCTYPE html>
<html lang="en">
<head>
<meta charset="UTF-8">
<meta name="viewport" content="width=device-width,
initial-scale=1">
href="https://area405.com/xmlrpc.php">
<title>Cite Black Authors: help us collect websites</title>
```

# Web ARChive (WARC) file format

```
WARC/1.0
WARC-Type: response
WARC-Target-URI: https://www.contemporary.org/
WARC-Date: 2023-11-25T23:35:55Z
WARC-Payload-Digest: sha1:MLCZXFBA3KL244XXDHEWQQKBY3FW6QEN
WARC-Record-ID: <urn:uuid:e629a74e-9824-4684-8831-5b8571443386>
Content-Type: application/http; msgtype=response
Content-Length: 36264

HTTP/1.1 200 OK
Server: cloudflare
Date: Sat, 25 Nov 2023 23:35:55 GMT
Content-Type: text/html; charset=UTF-8
Connection: close
Vary: Accept-Encoding
Last-Modified: Wed, 22 Jan 2022 02:14:20 GMT

<!DOCTYPE html>
<html lang="en">
<head>
<meta charset="UTF-8">
<meta name="viewport" content="width=device-width,
initial-scale=1">
href="https://contemporary.org/xmlrpc.php">
<title>The Contemporary, Baltimore, MD.</title>
```

**Response Record:**

Target URL

WARC Date & Time

Header

HTTP Status Code

Server Name

Content Type

Original payload to produce in a web browser

Content Block

# Web ARChive (WARC) file format

**References:**

- **Introduction to the WARC file** - Archive-It
- **WARC specification repository** - IIPC
- **WARC format description** - Library of Congress
- **Details for WARC 1.0** - PRONOM

# Web ARChive (WARC) file format

**Challenges:**

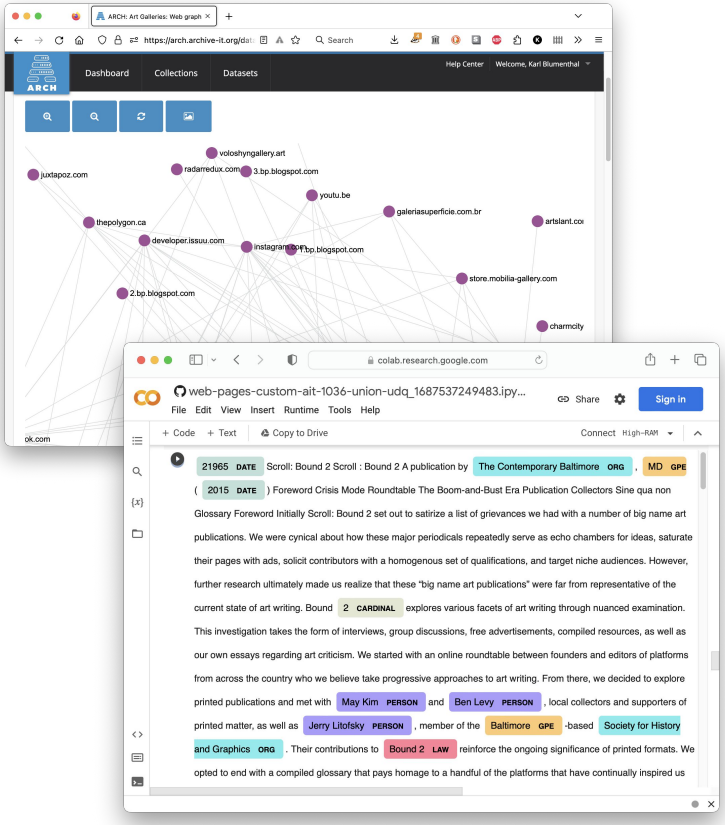
- Getting access to readily usable data can be labor intensive and expensive for GLAM practitioners and researchers.
- End users especially bear the burden of big, unwieldy datasets.
- Researchers need a community of practice around web archives to share methods, data, questions, and ideas.

# Archives Research Compute Hub (ARCH)

**Primary ARCH features:**

- Easily create machine actionable data from collections at scale
- Accessible interface designed with user community leadership
- In-browser visualization, exploratory analysis notebooks & tutorials
- Open source technology (Github repository)

# Archives Research Compute Hub (ARCH)


ARCH: Art Galleries: Web graph
Dashboard
Collections
Datasets
Help Center
Welcome, Karl Blumenthal
voloshyngallery.art
radarredux.com
3.bp.blogspot.com
juxtapoz.com
youtu.be
galeriasuperficie.com.br
thepolygon.ca
developer.issuu.com
1.bp.blogspot.com
store.mobilia-gallery.com
2.bp.blogspot.com
charmcity
colab.research.google.com
web-pages-custom-ait-1036-union-udq_1687537249483.ipy...
File Edit View Insert Runtime Tools Help
Share
Sign in
+ Code + Text
Copy to Drive
Connect High-RAM
21965 DATE Scroll: Bound 2 Scroll : Bound 2 A publication by The Contemporary Baltimore ORG , MD GPE
( 2015 DATE ) Foreword Crisis Mode Roundtable The Boom-and-Bust Era Publication Collectors Sine qua non
Glossary Foreword Initially Scroll: Bound 2 set out to satirize a list of grievances we had with a number of big name art
publications. We were cynical about how these major periodicals repeatedly serve as echo chambers for ideas, saturate
their pages with ads, solicit contributors with a homogenous set of qualifications, and target niche audiences. However,
further research ultimately made us realize that these "big name art publications" were far from representative of the
current state of art writing. Bound 2 CARDINAL explores various facets of art writing through nuanced examination.
This investigation takes the form of interviews, group discussions, free advertisements, compiled resources, as well as
our own essays regarding art criticism. We started with an online roundtable between founders and editors of platforms
from across the country who we believe take progressive approaches to art writing. From there, we decided to explore
printed publications and met with May Kim PERSON and Ben Levy PERSON , local collectors and supporters of
printed matter, as well as Jerry Litofsky PERSON , member of the Baltimore GPE -based Society for History
and Graphics ORG . Their contributions to Bound 2 LAW reinforce the ongoing significance of printed formats. We
opted to end with a compiled glossary that pays homage to a handful of the platforms that have continually inspired us

Non-profit infrastructure, open-source tools for computational use of digital collections.

Currently supports:

- Archive-It, Archive.org, and Vault
- Text, network, and binary analyses
- (Tr)OCR, NER, & speech-to-text

In development:

- Image classification & description
- Remote collection support

# ARCH demo

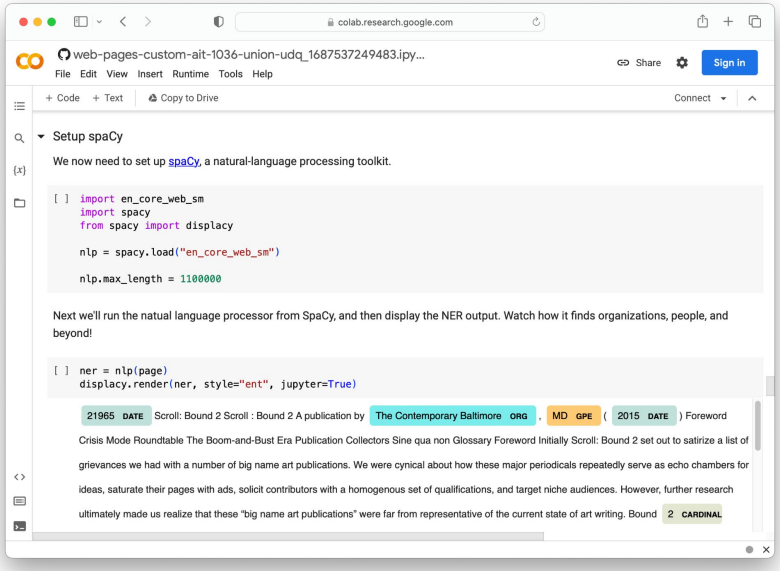
**https://arch.archive-it.org/**
u: **arlis**
p: **webinar2025**

ARLIS/NA | May 21, 2025

bit.ly/arlis-data ⬇️

# Methods and tools

**Tutorials:** Try ARCH datasets and tools on demand and at home!


colab.research.google.com
web-pages-custom-ait-1036-union-udq_1687537249483.ipy...
File Edit View Insert Runtime Tools Help
Share
Sign in
+ Code + Text Copy to Drive
Connect
Setup spaCy
We now need to set up spaCy, a natural-language processing toolkit.
[ ] import en_core_web_sm
import spacy
from spacy import displacy
nlp = spacy.load("en_core_web_sm")
nlp.max_length = 1100000
Next we'll run the natual language processor from SpaCy, and then display the NER output. Watch how it finds organizations, people, and beyond!
[ ] ner = nlp(page)
displacy.render(ner, style="ent", jupyter=True)
21965 DATE Scroll: Bound 2 Scroll : Bound 2 A publication by The Contemporary Baltimore ORG , MD GPE ( 2015 DATE ) Foreword
Crisis Mode Roundtable The Boom-and-Bust Era Publication Collectors Sine qua non Glossary Foreword Initially Scroll: Bound 2 set out to satirize a list of
grievances we had with a number of big name art publications. We were cynical about how these major periodicals repeatedly serve as echo chambers for
ideas, saturate their pages with ads, solicit contributors with a homogenous set of qualifications, and target niche audiences. However, further research
ultimately made us realize that these "big name art publications" were far from representative of the current state of art writing. Bound 2 CARDINAL

See our latest tutorials (and videos!) for:

- Text mining with Voyant
- Network analysis with Palladio and Gephi
- Image galleries with Palladio & Juxta
- Notebooks for CLI tools like nltk, spaCy, NER, and more

# Questions and discussion

**Share your thoughts:**

- What should CARTA collect and why?

# Questions and discussion

**Share your thoughts:**

- What is a research question you've encountered recently that you think could use a computational approach with web archive collection data?

# Questions and discussion

**Share your thoughts:**

- What barriers are there to computational research with collections at your institution?

# Thanks!

**Learn more and get started:**

- https://carta.archive-it.org/
- https://webservices.archive.org/pages/arch

**Karl Blumenthal**
Senior Web Archivist
Internet Archive
karlb@archive.org

**Sumitra Duncan**
Head, Web Archiving Program
NYARC
duncan@frick.org



bit.ly/arlis-data